Proceedings of the 2013 Winter Simulation Conference
R. Pasupathy, S.-H. Kim, A. Tolk, R. Hill, and M. E. Kuhl, eds.

DENSITY ESTIMATION OF SIMULATION OUTPUT USING EXPONENTIAL EPI-SPLINES

Dashi I. Singham
Johannes O. Royset

Department of Operations Research
Naval Postgraduate School
1411 Cunningham Road
Monterey, CA 93940, USA

Roger J-B Wets

Department of Mathematics
University of California, Davis
One Shields Avenue
Davis, CA 95616, USA

ABSTRACT

The density of stochastic simulation output provides more information on system performance than the mean alone. However, density estimation methods may require large sample sizes to achieve a certain accuracy or desired structural properties. A nonparametric estimation method based on exponential epi-splines has shown promise to overcome this difficulty by incorporating qualitative and quantitative information that reduces the space of possible density estimates substantially. Such 'soft' information may come in the form of the knowledge of a non-negative support, unimodality, and monotonicity, and is often available in simulation applications. We examine this method for output analysis of stochastic systems with fixed input parameters, and for a model with stochastic input parameters, with an emphasis on the use of derivative information.

1 INTRODUCTION

Simulation models are often analyzed to estimate the mean performance of a proposed system, with confidence intervals quantifying the uncertainty in the estimate. Quantile estimates of simulation output at various probability levels provide further information about the system performance (see Heidelberger and Lewis (1984) and Chen and Kelton (2006)). Estimates of the density of the output yield an even more comprehensive picture that allows the computation of these and many more quantities. While kernel methods and other traditional nonparametric density estimation methods can in principle be used to construct density estimates, they do not easily account for known structural properties and tend to require a large sample to reduce the effect of outliers. In contrast, parametric density estimation tends to impose excessive restrictions on the shape of the densities and important characteristics of the output thereby may remain undetected.

Royset and Wets (2013) introduce a nonparametric method for estimating density functions using exponential epi-splines that easily incorporates *soft information* about a stochastic system, its inputs, and its outputs. Soft information reduces the space of candidate density estimates substantially in many applications and may come in the form of knowledge of the nonnegativity of the support as well as monotonicity, unimodality, continuity, and smoothness of the density. For example, waiting times in queueing systems are nonnegative, and possibly bounded from above if entities renege after a certain amount of time. Under approximate normality, the density can reasonably be assumed to be unimodal. Exact and approximate density function values at particular points, cumulative distribution function values, and derivative information may also be available as we exemplify further below. In this paper, we apply the nonparametric method of Royset and Wets (2013) to estimate densities of simulation output and illustrate the use of soft information in this context. We construct density estimates to assess the variation in output from a stochastic simulation with fixed input parameters. For example, the input parameters may define

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
DEC 2013		**00-00-2013 to 00-00-2013**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Density Estimation of Simulation Output Using Expotential EPI-Splines	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **Naval Postgraduate School,Department of Operations Research,Monterey,CA,93943**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES
Proceedings of the 2013 Winter Simulation Conference, 8-11 Dec, Washington, DC, pp. 755-765.

14. ABSTRACT
The density of stochastic simulation output provides more information on system performance than the mean alone. However, density estimation methods may require large sample sizes to achieve a certain accuracy or desired structural properties. A nonparametric estimation method based on exponential epi-splines has shown promise to overcome this difficulty by incorporating qualitative and quantitative information that reduces the space of possible density estimates substantially. Such ?soft? information may come in the form of the knowledge of a non-negative support, unimodality, and monotonicity, and is often available in simulation applications. We examine this method for output analysis of stochastic systems with fixed input parameters, and for a model with stochastic input parameters, with an emphasis on the use of derivative information.

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**	**Same as Report (SAR)**	**11**	

the distribution of input random variates. However, these input parameters may be uncertain and varying them is also important for sensitivity analysis. While this is a situation which is typically computationally demanding, we demonstrate that when including soft information, such as knowledge of derivatives, high-quality estimates of the output density are achieved with as little as 10 runs of the simulation model. In principle, the nonparametric method of Royset and Wets (2013) applies in any finite dimension, but we here focus on a single output quantity of interest.

Exponential epi-splines are functions that are compositions of the exponential function with epi-splines. Epi-splines are piecewise polynomial as related to classical splines, but instead of focusing on interpolation aim at approximation. They also relax the typical requirement of continuity and smoothness of splines. Exponential epi-splines approximate to an arbitrary accuracy essentially any density function encountered in practice and also easily allow for the consideration of soft information. Royset and Wets (2013) develop a nonparametric maximum-likelihood based density estimator using exponential epi-splines and show asymptotic and finite sample-size results, including almost-sure convergence to the true density as the sample size tends to infinity. By restricting the shape of possible densities to consider, fewer samples are often required to construct an estimate that meets the requirements of the analyst. Simulation experiments are usually costly and time consuming to run, and therefore this estimator appears especially well suited to that context. The use of exponential epi-spline estimators of output densities differs substantially from current analysis of simulation output; see Alexopolous (2006) for a general overview of simulation output analysis and Chen and Kelton (2008) for density estimation methods using histograms.

The remainder of the paper is organized as follows. Section 2 provides background on the nonparameteric method of Royset and Wets (2013). Section 3 discusses different types of soft information. Section 4 presents computational experiments, and Section 5 concludes and discusses future applications.

2 EPI-SPLINES FOR DENSITY ESTIMATION

We briefly outline the density estimation method developed in Royset and Wets (2013) and let x be a random variable, with density h, describing an output quantity of interest from a stochastic simulation. An exponential epi-spline estimate of h, based on a sample of size n, is given by

$$h^n(x) = e^{-s^n(x)}, \quad x \in \mathbb{R}, \tag{1}$$

where s^n is an epi-spline constructed from the sample as described below. We start, however, with a description of epi-splines.

In general, epi-splines are real-valued functions defined on a closed interval that is partitioned by a mesh $m = (m_0, m_1, \ldots, m_N)$, with $m_{k-1} < m_k$ for $k = 1, 2, \ldots, N$. An epi-spline is of order p if $s : [m_0, m_N] \to \mathbb{R}$ is polynomial of degree p in each open segment $(m_{k-1}, m_k), k = 1, 2, \ldots, N$, and is finite valued at the mesh points $m_0, m_1, \ldots, m_N$. Exponential epi-splines of order two capture many common densities exactly, including the normal and exponential densities, and essentially all other densities to an arbitrary accuracy as the number of partitions N grows; see Royset and Wets (2013) for details.

An epi-spline s of order p and mesh $m = (m_0, m_1, ..., m_N)$ is defined by $(p+2)N + 1$ parameters, which allow for finite-parameter optimization in a maximum-likelihood estimation problem. Specifically, we define the epi-spline parameter $r = (s_0, s_1, \ldots, s_N, a_1, a_2, \ldots, a_N)$, where $(s_0, s_1, \ldots, s_N)$ are the values of the epi-spline s on the mesh. The remaining parameters $(a_1, a_2, \ldots a_N)$ are each $(p+1)$-tuples that contain the polynomial coefficients for the corresponding interval. This means that for a given epi-spline s and a particular value x in an interval (m_{k-1}, m_k), we have

$$s(x) = \sum_{i=0}^{p} a_{k,i}(x - m_{k-1})^i.$$

Combining these facts, we find that $s(x) = \langle c(x), r\rangle$ for all $x \in [m_0, m_N]$, with $c : [m_0, m_N] \to \mathbb{R}^{(p+2)N+1}$ being defined by

$$c(x) = \begin{cases} (0_{N+1+(p+1)(k-1)}, 1, (x-m_{k-1}), \ldots, (x-m_{k-1})^p, 0_{(p+1)(N-k)}), \\ \qquad\qquad \text{if } x \in (m_{k-1}, m_k), k = 1, 2, \ldots, N; \\ (0_k, 1, 0_{N-k+(p+1)N}), \qquad \text{if } x = m_k, k = 0, 1, \ldots, N. \end{cases}$$

Consequently, for a given mesh, an epi-spline of order p is uniquely defined by its epi-spline parameter r as the 'basis' function c is simply determined by the mesh and p.

Returning to (1), we obtain s^n, or equivalently the corresponding epi-spline parameter r^n, by maximizing the likelihood that the corresponding exponential epi-spline $\exp(-s^n(\cdot)) = \exp(-\langle c(\cdot), r^n\rangle)$ is the density corresponding to a sample. Let $x_1, \ldots, x_n$ be a sample realization. Then, maximizing the likelihood is equivalent to minimizing the negative log-likelihood, which yields the problem

$$\begin{aligned} \mathbf{MLP}^n : \quad \min_r \frac{1}{n}\sum_{i=1}^{n}\langle c(x_i), r\rangle & \\ \text{s.t.} \int_{d_0}^{d_N} e^{-\langle c(x), r\rangle} dx &= 1; \\ r &\in R^n, \end{aligned}$$

where we take advantage of the fact that the optimization can be carried out over the epi-spline parameter r, which consists of $(p+2)N+1$ variables. Moreover, R^n is a set, typically polyhedral or at least convex, defined by soft information as exemplified in Section 3. Since the objective function in $\mathbf{MLP}^n$ is linear, the main challenge with the problem is therefore the nonconvex integral constraint. However, under broad conditions this constraint can be relaxed to a convex constraint resulting in a convex optimization problem solvable by many efficient algorithms; see Royset and Wets (2013) for details. Even in nonconvex cases, our computational experience over thousands of instances indicates that high-quality solutions are easily obtained by standard solvers. An optimal solution, denoted by r^n, results in an epi-spline $s^n(\cdot) = \langle c(\cdot), r^n\rangle$ and an exponential epi-spline estimate $h^n(\cdot) = \exp(-s^n(\cdot)) = \exp(-\langle c(\cdot), r^n\rangle)$. Under mild assumptions, as $n \to \infty$, h^n converges uniformly to h (except possibly at the mesh points), and the sequence of solutions r^n converges to the epi-parameter of the true density; see Royset and Wets (2013) for details.

The above nonparametric density estimation method applies broadly, but the simulation context poses special challenges. For example, suppose that we have a stable single-server queueing simulation where the average customer time in system (TIS) is a function of the mean service time parameter ω with the arrival parameters fixed. The output x is the steady-state average customer TIS. We seek the density of x under uncertainty in ω given by some distribution. In principle, one way to estimate this density is to sample values of ω, simulate the corresponding values of x, and compute an exponential epi-spline estimate of the true density. However, since each value of ω requires, most likely, a long simulation to obtain x, the computational cost may become high and effective estimates from small sample sizes become especially pertinent. In addition, the resulting output will be corrupted by simulation error. As an illustration, if the support of the density of ω is bounded, then the corresponding values of x should also be bounded, but sampling error may lead to values of x outside those bounds. This sampling error can be managed, and it suffices for the strong consistency of the exponential epi-spline estimator in Royset and Wets (2013) that the probability measure generated by a sample tends to the true measure in a probability metric almost surely. However, we do not address this in further detail, but include a numerical example in Section 4.

3 SOFT INFORMATION

The soft information defining the region R^n comes in a variety of forms and here we simply give specific examples. We start, however, by observing that the mesh $m = (m_0, m_1, ..., m_N)$ is typically also informed

by soft information. If the system output is bounded from above and below, then m_0 and m_N are naturally given by those bounds. The spacing between mesh points can be uniform, or with more closely spaced points in regions where the density is known to vary the most.

We recall the epi-spline parameter

$$r = (s_0, s_1, \ldots, s_N, a_1, a_2, \ldots, a_N),$$

with $a_k = (a_{k,0}, a_{k,1}, ..., a_{k,p})$ for all $k = 1, 2, ..., N$. Using these parameters, we formulate the following constraints:

Continuity. We ensure that an exponential epi-spline estimate is lower semi-continuous by imposing the constraint

$$s_{k-1} \geq a_{k,0}, \quad s_k \geq \sum_{i=0}^{p} a_{k,i}(m_k - m_{k-1})^i, \quad k = 1, 2, ..., N.$$

Continuity would require the same constraints with inequalities replaced by equalities.

Smoothness. We ensure a continuously differentiable density by imposing the conditions for continuity and

$$\sum_{i=1}^{p} i a_{k,i}(m_k - m_{k-1})^{i-1} = a_{k+1,1}, \ k = 1, 2, ..., N-1.$$

Pointwise Fisher information. We define the pointwise Fisher information of an exponential epi-spline density h at x to be

$$h'(x)/h(x) = -\langle c'(x), r\rangle = -\sum_{i=1}^{p} i a_{k,i}(x - m_{k-1})^{i-1}$$

for $x \in (m_{k-1}, m_k)$, which can then be bounded from above and/or below.

Monotonicity. We achieve a nondecreasing (nonincreasing) density by imposing nonnegativity (non-positivity) on $h'(x)/h(x)$ for all $x \in (m_{k-1}, m_k), k = 1, 2, ..., N$ as described above, as well as for $k = 1, 2, ..., N$,

$$s_{k-1} \geq (\leq) a_{k,0}, \quad s_k \leq (\geq) \sum_{i=0}^{p} a_{k,i}(m_k - m_{k-1})^i.$$

Strongly unimodal. A density is strongly unimodal if it is continuous and log-concave. Consequently, if $h(x) = \exp(-\langle c(x), r\rangle)$, $x \in [m_0, m_N]$, strong unimodality requires that $\langle c(x), r\rangle$ is a convex function in x. This condition is ensured if (i) continuity is imposed (see above), (ii) for $k = 1, 2, ..., N-1$, the epi-spline's left derivative at m_k is no larger than its right derivative, i.e., for $k = 1, 2, ..., N-1$,

$$\sum_{i=1}^{p} i a_{k,i}(m_k - m_{k-1})^{i-1} \leq a_{k+1,1},$$

and (iii) on each (m_{k-1}, m_k), $k = 1, 2, ..., N$, $\langle c(x), r\rangle$ is a convex function in x, i.e., for $k = 1, 2, ..., N$, $x \in (m_{k-1}, m_k)$,

$$\sum_{i=2}^{p} i(i-1) a_{k,i}(x - m_{k-1})^{i-2} \geq 0.$$

Here, the obvious interpretations are required when $p = 0, 1$.

Bounds on cumulative distribution functions. If the cumulative distribution function at $\gamma \in [m_0, m_N]$ of the estimated density h^n must lie between the lower bound l and the upper bound u, then we obtain the following constraints

$$\int_{m_0}^{\gamma} e^{-\langle c(x),r\rangle} dx \leq u \text{ and } \int_{\gamma}^{m_N} e^{-\langle c(x),r\rangle} dx \leq 1 - l.$$

Bijective functions and derivative information. We next discuss derivative information about the stochastic system and suppose that the output quantity of interest is (conceptually) given by

$$x = g(\omega),$$

where g is a real-valued function that takes as input ω. The stochasticity of the system derives from the uncertainty about ω, which may be modeled by a random vector or stochastic process in general. In this paragraph, however, we assume that ω is a scalar-valued random variable. (While most stochastic simulations involve more than one input parameter, we here envision, for example, a study of a single input parameter of high importance, with other parameters fixed.) We assume that for a given ω, the simulation returns $g(\omega)$ and the derivative $g'(\omega)$. We refer to Fu (2006) for methods to generate derivatives such as infinitesimal perturbation analysis. If the function g is differentiable and bijective, then derivative information can be used to calculate the output density exactly at certain points. Specifically, if the derivative of g and the density of ω (denoted as $f_\omega(\omega)$) are known, the density of the output can be calculated exactly at certain points using the following change of variables formula:

$$h(x) = h(g(\omega)) = \frac{1}{|g'(\omega)|} f_\omega(\omega).$$

This information results in the following constraints for the exponential epi-spline estimate at any x for which the true density h is known

$$\sum_{i=0}^{p} a_{k,i}(x - m_{k-1})^i = -\log h(x) \tag{2}$$

where k is the interval such that $x \in (m_{k-1}, m_k)$. Of course, given a specific simulation model, derivatives of g or even the value x may not be available exactly and the above equality may be replaced by

$$-\log \tilde{h}(\tilde{x}) - \epsilon \leq \sum_{i=0}^{p} a_{k,i}(x - m_{k-1})^i \leq -\log \tilde{h}(\tilde{x}) + \epsilon$$

where $\tilde{h}$ and $\tilde{x}$ are approximate values, and ϵ a tolerance.

4 COMPUTATIONAL EXAMPLES

We describe three examples that illustrate the method. We choose examples where the true density is known so that a mean squared error is computable as

$$\text{MSE} = \int (h^n(x) - h(x))^2 h(x) dx.$$

All results are computed using common random numbers for 100 replications, with different levels of soft information applied. We compare exponential epi-spline estimates of order two to those calculated using kernel estimation. We adopt the standard kernel density estimator in MATLAB version R2012a with a gaussian kernel, and do not modify the default (optimized) calculation of the bandwidth. **MLP**^{n} is solved using fmincon of MATLAB, typically in under one minute of run time on a dual-core 2.2 GHz processor.

4.1 Lognormal Distribution

As a basic example to illustrate the benefits of using soft information, we consider the case where ω is a standard normal random variable and $x = e^{\omega}$, so x has a standard lognormal distribution. We estimate the density for 100 replications, each with samples of 10 points. Table 1 shows the average and standard deviation of MSE values across the 100 replications. We perform eight experiments using exponential epi-spline estimates with different combinations of soft information. As with any non-parametric estimation, some constraints on the class of functions under consideration are required to obtain meaningful results. All exponential epi-splines are computed under the assumptions of continuity, smoothness, pointwise Fisher information in $[-2, 2]$, and number of partitions $N = 20$. The mesh partitions are determined by taking equally spaced intervals within the range of sample data widened by one standard deviation at each end.

Table 1: MSE results for lognormal data.

Information	Average MSE	Standard Deviation
Unimodal (U)	0.0385	0.0251
Lower Bound (LB)	0.0254	0.0194
Median (CDF)	0.0340	0.0214
U, LB, CDF	0.0267	0.0233
U, LB, CDF and Derivative (1pt)	0.0245	0.0190
U, LB, CDF and Derivative (2pts)	0.0183	0.0185
U, LB, CDF and Derivative (5pts)	0.0060	0.0071
U, LB, CDF and Derivative (10pts)	0.0036	0.0035
Kernel	0.0416	0.0373

The first row constrains the density to be unimodal, and the second places a lower bound on the support at zero. Because we know that the true median is 1, we can also constrain the cumulative distribution function to be 0.5 at 1 as a third experiment. A fourth experiment includes all three classes of constraints. In view of Table 1, it appears that adding the lower bound contributes the greatest reduction in MSE out of the first four experiments. Rows 5-8 show the effect from constraining the density values at observed realizations of x using (2). Constraining the density at more points leads to better estimates as expected, and increasing the amount of soft information also generally results in decreasing the standard deviations of the estimates. We find that these types of derivative constraints contribute to a major reduction in MSE, especially when combined with other soft information. All of the experiments have lower average MSE values than the kernel estimate.

Figure 1 shows typical results for the exponential epi-spline and kernel estimates. In addition to having lower MSE values, exponential epi-splines meet qualitative requirements and visually tend to match the true density quite well. In particular, adding derivative information improves 'closeness' of the exponential epi-spline estimate to the true density dramatically. Without ad hoc modifications, kernel estimates might deliver positive density values for negative x (violating the lower bound constraint) and have bumps at outliers (violating the unimodality constraint).

4.2 M/M/1 Queueing Example

This example illustrates the effect of imprecise data from realizations of the output x obtained by simulation. We return to the example in the last paragraph of Section 2 and aim to estimate the density of the mean customer TIS given that there is uncertainty in the mean service rate parameter. Consider the M/M/1 queue with λ as the arrival rate and ω as the service rate with $\lambda < \omega$. Then, the mean TIS is $(\omega - \lambda)^{-1}$.

We assume that $\lambda = 1$, but there is uncertainty in ω, which has a uniform distribution $U(1.5, 2.5)$; see Lim and Glynn (2006) and Staum (2009) for similar examples. The mean TIS, x, of the customers is now a function of ω. Specifically, $x = g(\omega) = (\omega - 1)^{-1}$, with derivative $g'(\omega) = -(\omega - 1)^{-2}$. We want

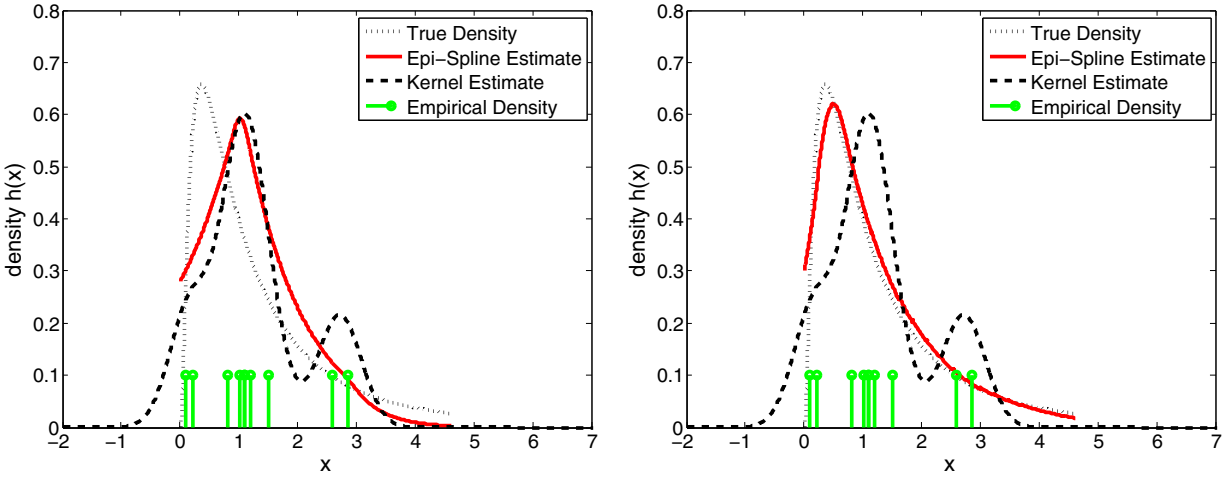

Figure 1: Example density estimates for lognormal density data. Left: Lower bound at zero and unimodality constraints are included as soft information (the MSE is 0.0261). Right: Added derivative information for 10 points (the MSE reduces to 0.0020). The unmodified kernel estimate MSE is 0.0411.

to estimate the density of x given the uncertainty in ω. Even though g and g' are known exactly for the M/M/1 queue, we simulate these quantities for each ω using one run of 1 million customers. Of course, in more general systems, simulation would be the only option to generate sample points.

Table 2: MSE results for the M/M/1 example.

Information	Average MSE	Standard Deviation
Unimodal (U)	0.4589	0.1577
Lower & Upper Bound (LU)	0.0799	0.0762
Decreasing (D)	0.4634	0.1472
U, LU, D	0.2285	0.1994
LU and Derivative (1pt)	0.0401	0.0543
LU and Derivative (2pt)	0.0220	0.0277
LU and Derivative (5pts)	0.0085	0.0139
LU and Derivative (10pts)	0.0011	0.0016
Kernel	0.4310	0.3536

In the exponential epi-spline estimates, we include continuity, differentiability, and pointwise Fisher information constraints with values in [-4,0], as well as $N = 20$. Each replication uses samples of 10 different ω values, each with a simulated mean TIS. Table 2 shows the results for a range of soft information. If we know the density of waiting times has a shape similar to an exponential distribution, we can say that the density is unimodal. Because the density of ω is bounded, we know that the true bounds of x are $[\frac{2}{3}, 2]$. However, we heuristically add an extra 0.005 to each end of the range to allow for sampling error. For this example, we may also know that the density function is decreasing. Knowledge of the lower and upper bounds greatly contributes to a reduction in MSE compared to the other types of soft information. Other experiments not shown here reveal that using more conservative lower and upper bounds lead to higher MSE values because epi-splines predict positive densities over areas outside the true range.

Finally, adding derivative information contributes to an additional reduction in MSE, as seen in Table 2. Even though x and its gradient information are estimated (so technically the soft information constraints are inexact), we still see a great reduction in MSE. The standard deviation also decreases with increasing soft information. Figure 2 shows an example of one replication. The left plot shows an epi-spline with

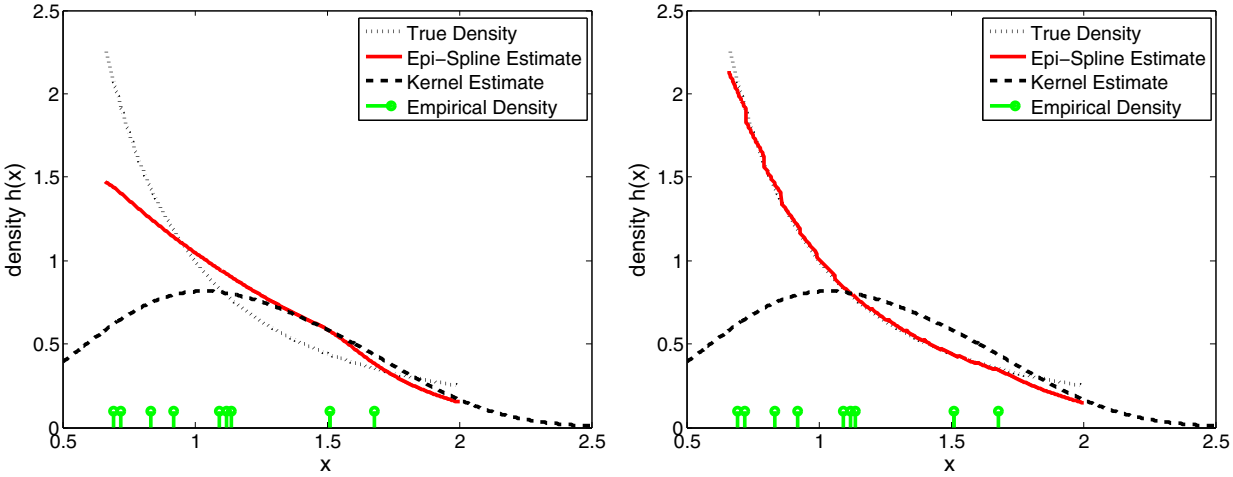

Figure 2: Example for M/M/1 queue mean TIS density. Left: Unimodal, lower & upper bound, & decreasing constraints (the MSE is 0.0675). Right: Added 10 points of derivative information (the MSE is 0.0013). The unmodified kernel estimate MSE is 0.4751.

the soft information constraints of unimodality, lower and upper bounds, and decreasing, and appears to be qualitatively closer than the kernel estimate. The right plot shows the inclusion of gradient information to constrain the density at the sampled values of x. We see that the epi-spline matches the true density closely in the regions where there are sample points, but not as closely between 1.75 and 2, where there are no sample points.

4.3 Simulated Queueing Model with Discontinuous Density

We now analyze output from a slightly more complex simulation model of a queueing system where the output density function is discontinuous. This system is an M/M/1 queue with arrival rate $\lambda = 1$ and service rate $\mu = 1.5$, but 50% of customers who enter the system are held at a separate station for two time units before entering the queue. We wish to estimate the density of the customer TIS. The true density is an equal mixture of the density of the M/M/1 TIS density, and that same density shifted to the right by two minutes. The resulting density is discontinuous at two as half the customers will have two minutes added to their time in system and is given by

$$h(x) = \frac{1}{2}(\mu - \lambda)\left(e^{-(\mu-\lambda)x} + e^{-(\mu-\lambda)(x-2)}\mathbb{I}_{x>2}\right)$$

where $\mathbb{I}_{x>2}$ is the indicator function for the event $x > 2$.

We use an event graph model with ranked lists (Schruben and Schruben 2009) to simulate and record tens of thousands of customer TIS values for this queueing system. Then, we sample 100 TIS values and compare the performance of epi-spline and kernel methods in terms of MSE. All experiments use a mesh $N = 10$. As exponential epi-spline estimates are not continuous (unless constrained to be so), they can be used to obtain reasonable estimates of a discontinuous density function when the mesh is fine enough to capture discontinuities.

Table 3 shows the effect of soft information, and we see that all the experiments perform similarly in terms of MSE. The first row shows the results with no information, and the second row assumes lower semicontinuity. The third row assumes continuity and differentiability (which we know to be false) and we see that the effect on the average MSE does not change much, though the standard deviation decreases. We can combine lower semicontinuity with a lower bound or Fisher information to obtain slightly improved

results. Finally, we add the constraint that the function is unimodal after a certain point, because the right tail of the true density is decreasing.

Table 3: MSE results for discontinuous density example.

Information	Average MSE	Standard Deviation
No information	0.0052	0.0070
Lower Semicontinuous (LSC)	0.0045	0.0020
Continuous & Differentiable	0.0045	0.0011
LSC and Fisher information (FI)	0.0040	0.0016
LSC and Lower Bound (LB)	0.0040	0.0021
LB, Unimodal Right Tail (URT)	0.0035	0.0019
LSC, FI, LB, and URT	0.0030	0.0017
Kernel	0.0042	0.0008

Figure 3 shows one replication, where the exponential epi-spline estimate better captures the discontinuity and shape of the density than the kernel estimate. The left plot shows the epi-spline with a lower bound, lower semicontinuous, and Fisher information constraints (within $[-4, 0]$), while the right plot added the constraint that the density must be unimodal in the right tail of the distribution, which greatly improves the appearance of the density. This example shows that epi-splines can provide a more reasonable structure for density estimates, even when MSE values are comparable to those of kernel methods. Varying the levels of soft information can result in densities with better qualitative properties, even if the MSE does not decrease greatly.

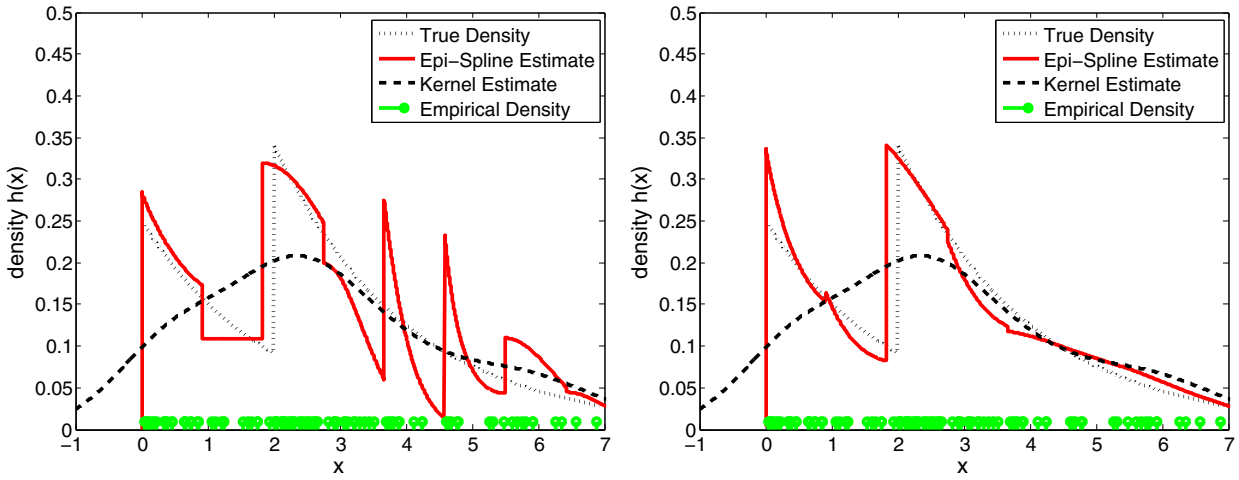

Figure 3: Example for discontinuous density function. Left: Lower semicontinuous, lower bound, and Fisher information (the MSE is 0.0020). Right: Added unimodality constraint to right tail (the MSE is 0.0016). The unmodified kernel estimate MSE is 0.0037.

5 CONCLUSIONS AND OTHER APPLICATIONS

Epi-splines with soft information can be used to construct density functions from limited data samples. Knowledge of soft information constrains the feasible space of possible epi-splines to provide better estimates. We show how simulation analysis can uniquely benefit from the availability of soft information, as the shape or bounds of the density functions for simulation output are often known. Simulation runs

can also be expensive, and soft information can help reduce the number of data points needed to construct an estimate. When derivative information is available, the reduction in MSE can be even greater.

Other more explicit information about the density might be known. If we have a reference density that we know will be close to the true density, we can place bounds on the divergence between the estimated density and this reference density. This may be useful in sensitivity analysis, where we have a good estimate for the density for a particular set of input parameters and wish to estimate the density for a slightly different set of parameters. In order to calculate this new density quickly (with a much smaller sample size than that used to calculate the reference density) we can bound the divergence from the old density to be small.

It should be noted that epi-splines can also be used to construct density functions from small data samples to generate input to a simulation model. The presence of soft information can be used to construct epi-spline estimates with more flexibility in shape than a parametric distribution. Additionally, simulation from epi-splines can be performed using acceptance/rejection techniques, where the majorizing function is a piecewise constant function which takes the maximum value of the epi-spline over each segment. Epi-spline density function values are easy to evaluate given that they are piecewise polynomial. We anticipate many future related applications for epi-splines.

ACKNOWLEDGMENTS

This material is based upon work supported in part by the U.S. Army Research Laboratory and the U.S. Army Research Office under grant numbers 00101-80683, W911NF-10-1-0246 and W911NF-12-1-0273.

REFERENCES

Alexopolous, C. 2006. "Statistical Estimation in Computer Simulation". In *Handbooks in Operations Research and Management Science: Simulation*, edited by S. Henderson and B. Nelson, Chapter 8, 193–223. Elsevier.

Chen, E., and W. Kelton. 2006. "Quantile and Tolerance-Interval Estimation in Simulation". *European Journal of Operational Research* 168 (2): 520–540.

Chen, E., and W. Kelton. 2008. "Estimating Steady-State Distributions via Simulation-Generated Histograms". *Computers & Operations Research* 35 (4): 1003–1016.

Fu, M. 2006. "Stochastic Gradient Estimation". In *Handbook on Operations Research and Management Science: Simulation*, edited by S. Henderson and B. Nelson, Chapter 19, 575–616. Elsevier.

Heidelberger, P., and P. Lewis. 1984. "Quantile Estimation in Dependent Sequences". *Operations Research* 32 (1): 185–209.

Lim, E., and P. W. Glynn. 2006. "Simulation-Based Response Surface Computation in the Presence of Monotonicity". In *Proceedings of the 2006 Winter Simulation Conference*, edited by L. Perrone, F. Wieland, J. Liu, B. Lawson, D. Nicol, and R. Fujimoto, 264–271. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Royset, J.O. and R. J-B. Wets 2013. "Nonparametric Density Estimation with Soft Information Using Exponential Epi-Splines". Under Review.

Schruben, D.L. and L.W. Schruben 2009. "Simulating Dynamic Systems with Event Relationship Graphs". http://www.CustomSimulations.com.

Staum, J. 2009. "Better Simulation Metamodeling: The Why, What, and How of Stochastic Kriging". In *Proceedings of the 2009 Winter Simulation Conference*, edited by M. D. Rossetti, R. R. Hill, B. Johansson, A. Dunkin, and R. G. Ingalls, 119–133. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

AUTHOR BIOGRAPHIES

DASHI SINGHAM is a Research Assistant Professor of Operations Research at the Naval Postgraduate School. Her research interests include simulation input and output analysis and applied statistics.

Application areas include energy, healthcare, and military modeling. She is an associate editor at *IIE Transactions*. Dashi has a bachelors degree in Operations Research & Financial Engineering from Princeton University, a Masters in Statistics and a Ph.D. in Industrial Engineering & Operations Research from the University of California at Berkeley (2010). Her email address is dsingham@nps.edu and her web page is http://faculty.nps.edu/dsingham.

JOHANNES O. ROYSET is an Associate Professor of Operations Research at the Naval Postgraduate School. His research focuses on formulating and solving stochastic and deterministic optimization problems arising in statistical estimation, complex systems, and sensor allocation, and has resulted in more than 50 technical publications. Dr. Royset has a Doctor of Philosophy degree from the University of California at Berkeley (2002). He was awarded a National Research Council postdoctoral fellowship in 2003, a Young Investigator Award from the Air Force Office of Scientific Research in 2007, and the Barchi Prize as well as the MOR Journal Award from the Military Operations Research Society in 2009. He received the Carl E. and Jessie W. Menneken Faculty Award for Excellence in Scientific Research in 2010. Dr. Royset is an associate editor of *Operations Research, Naval Research Logistics, Journal of Optimization Theory and Applications*, and *Computational Optimization and Applications*. His email address is joroyset@nps.edu.

ROGER J-B WETS is Distinguished Research Professor of Mathematics at the University of California, Davis. His research centers on stochastic optimization; equilibrium and optimization, especially in an uncertain environment; variational analysis; statistical estimation, in particular with poor and extremely poor data information; and mathematical finance. He has published about 200 technical articles. Dr. Wets received a Ph.D in Engineering Sciences from the University of California, Berkeley. His email address is rjbwets@ucdavis.edu.